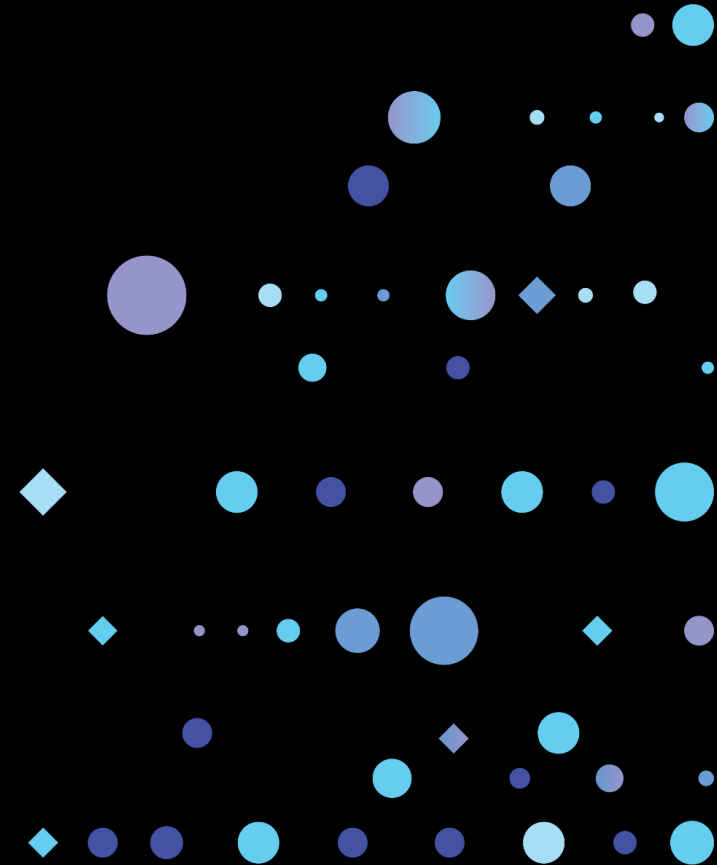




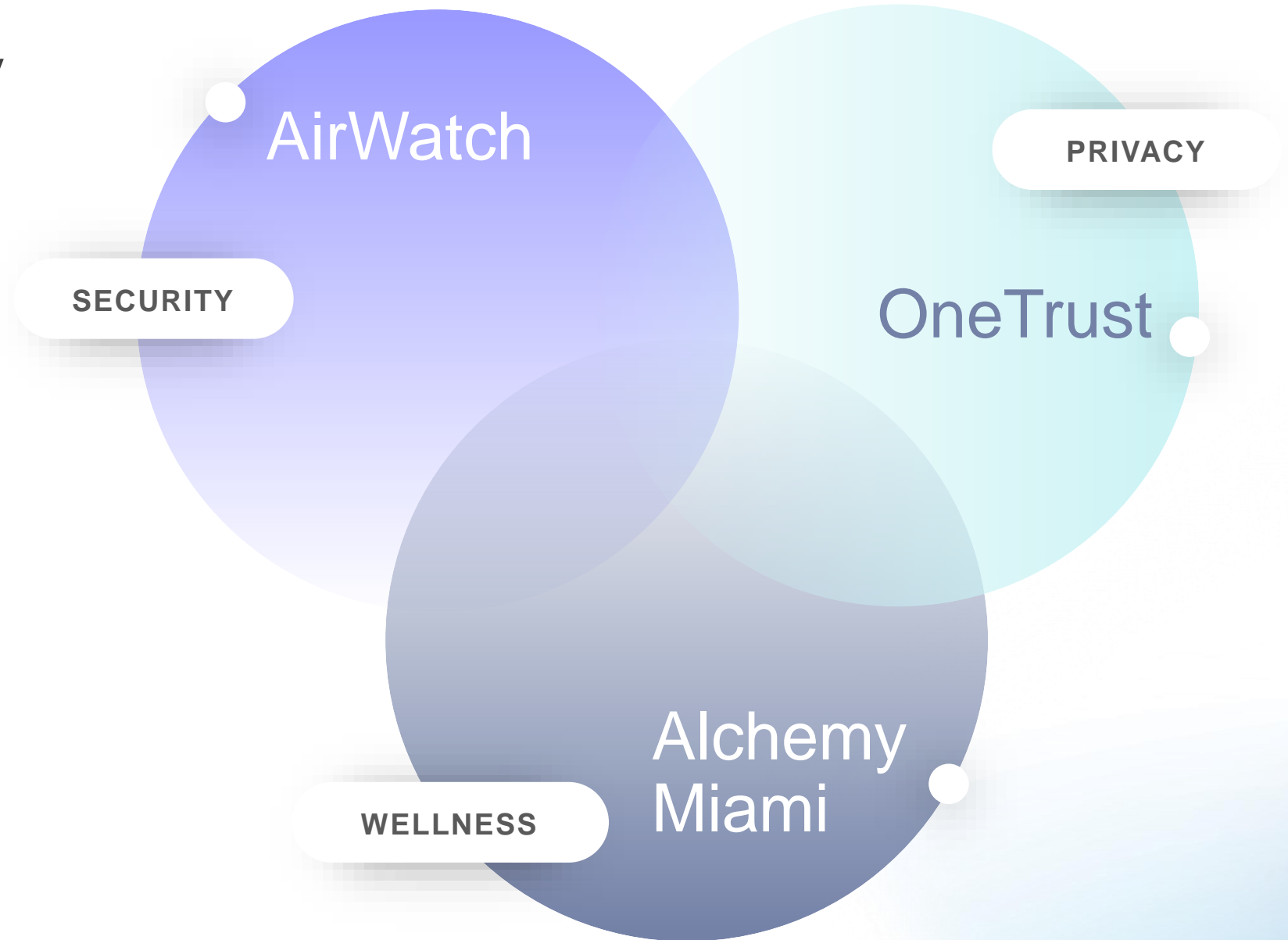
Innovate with AI

Enterprise AI Simplified

SEPTEMBER 11, 2024

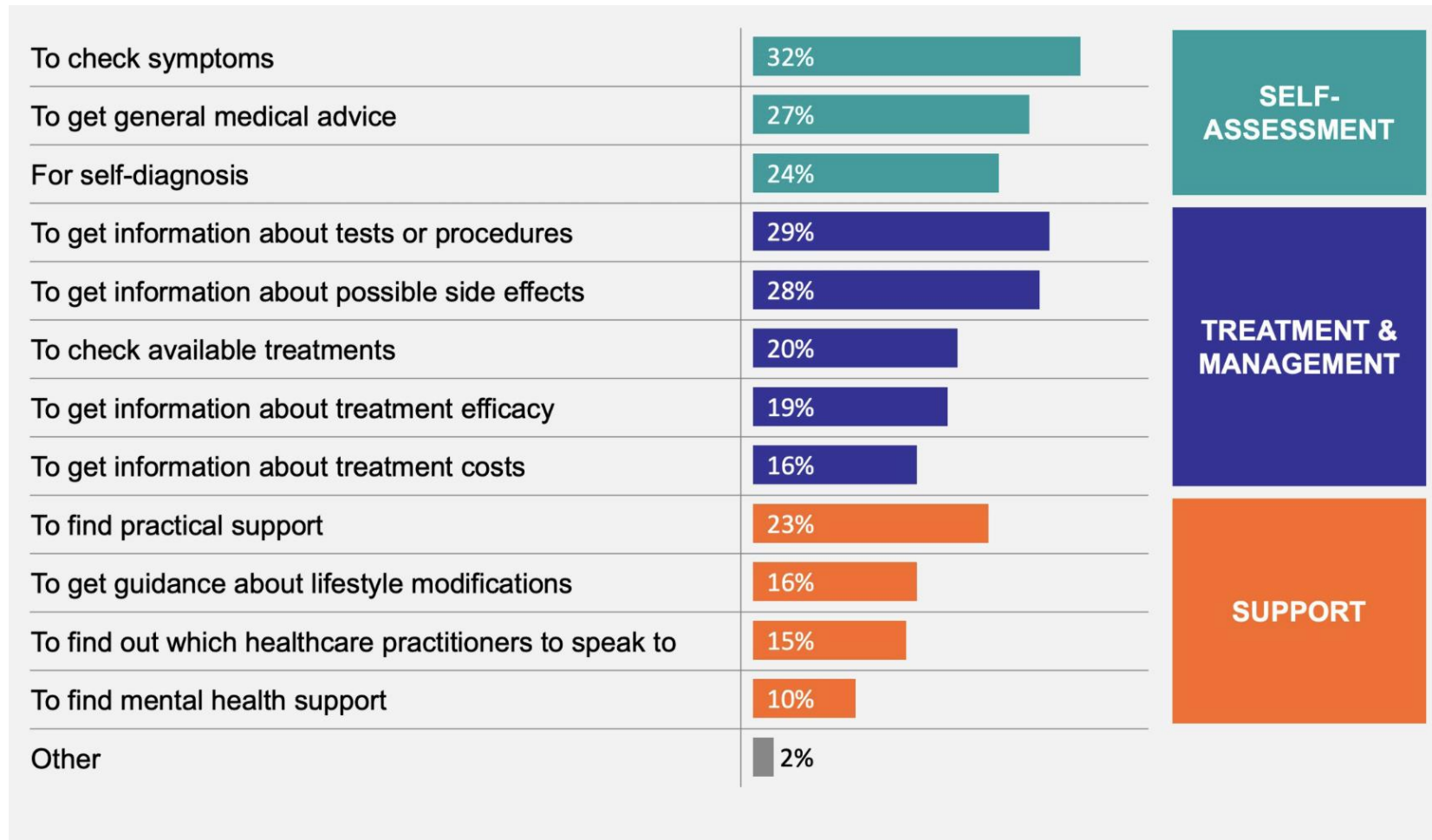


My AI Journey



Objectives of Consumers' Use of Generative AI

% surveyed consumers citing use of Generative AI



Create & Maximize AI Value

ADOPT EXISTING

MODEL

DATA

PROMPTS

GOVERNANCE

SECURITY



LOWER VALUE
HIGHER RISK



Effort & Complexity



HIGHER VALUE
LOWER RISK

CREATE & CONTROL

MULTIPLE MODELS

DOMAIN MODELS

COMPANY MODELS

TOOLS & TECHNOLOGIES

DEPLOYMENT STRATEGIES

GOVERNANCE POLICIES

SECURITY FRAMEWORKS

PROMPTS & FINE-TUNING

Speed Bumps to Realize Higher Value & ROI



Implementation / Data Integration



Legal and IP



Security



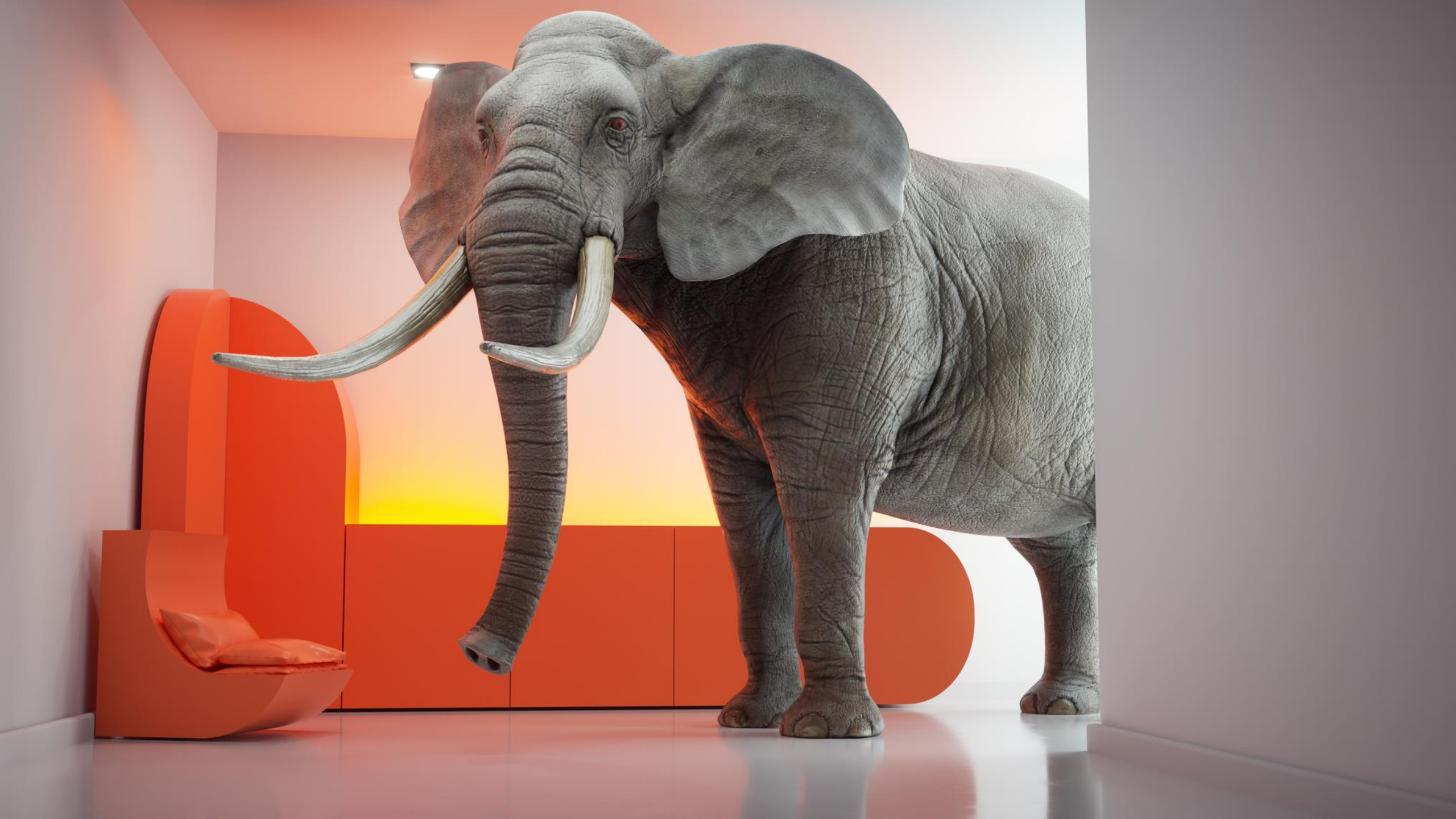
Regulatory and Governance Challenges



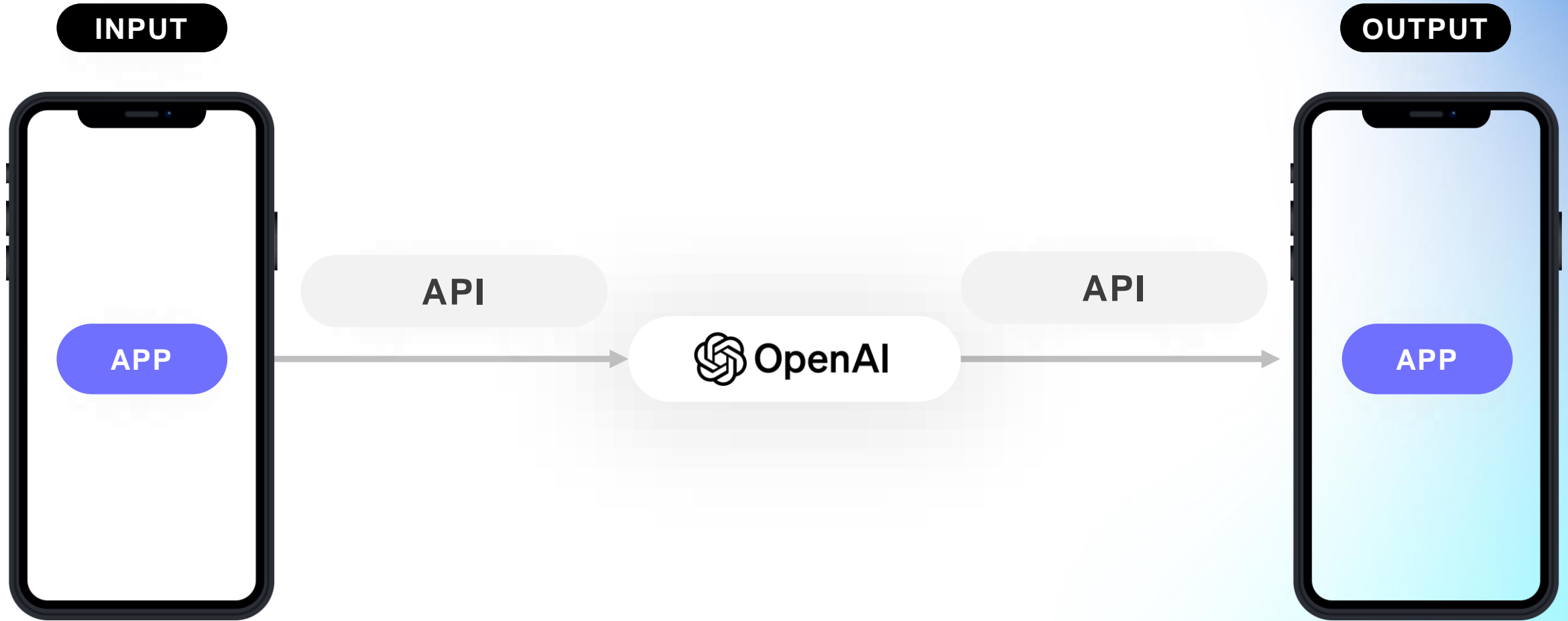
Cost



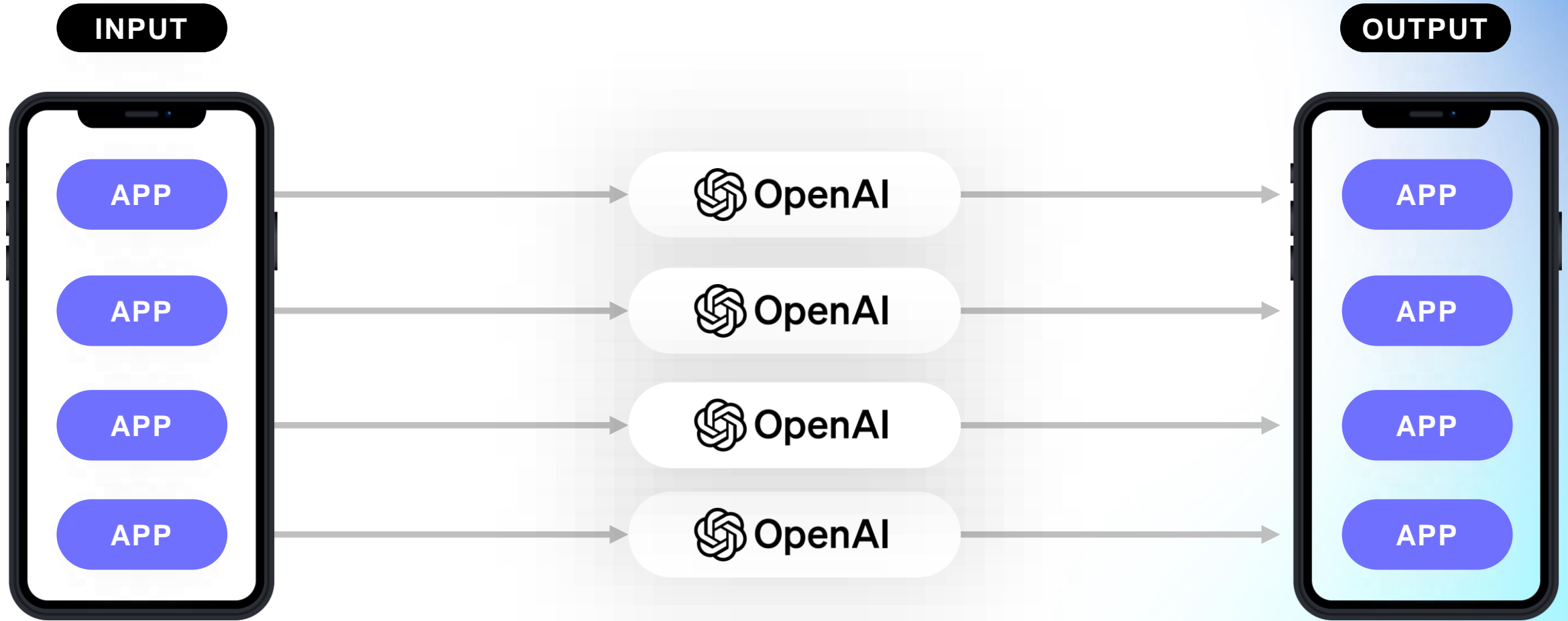
Brand and Reputational Risk



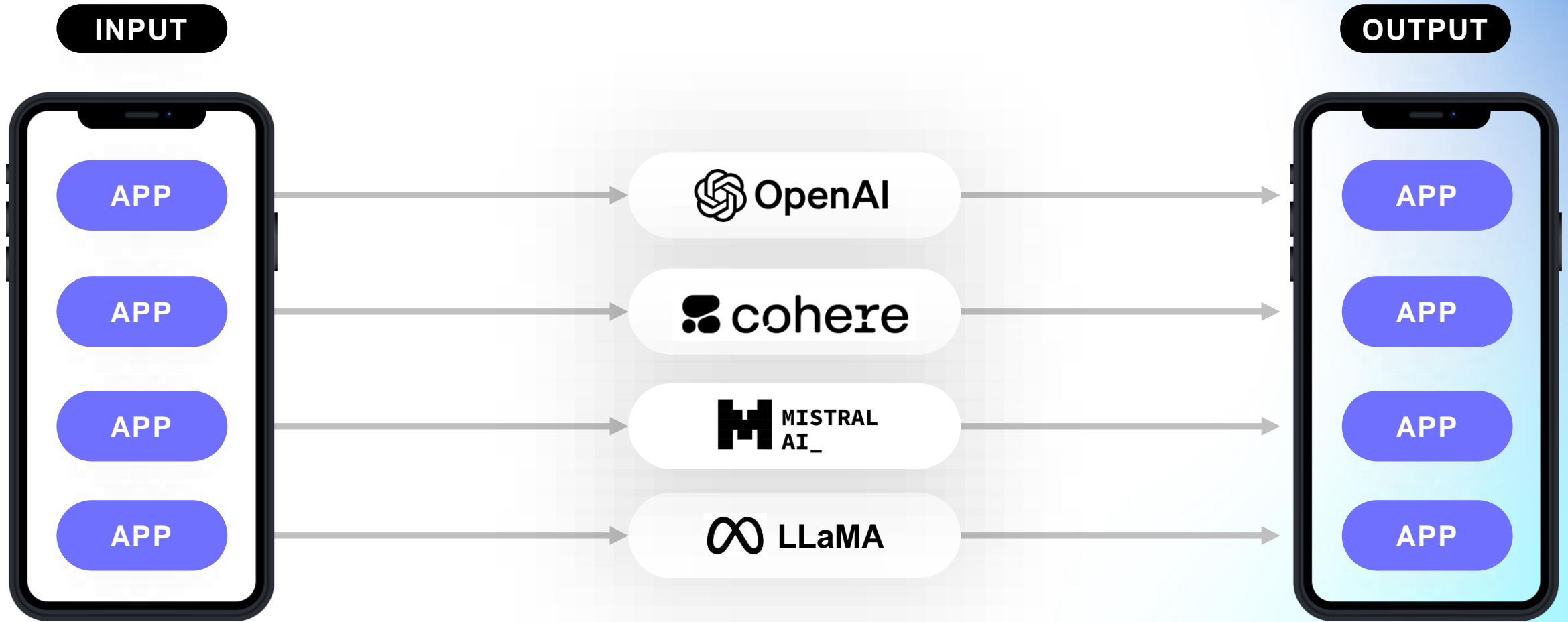
A Common Starting Point



Use Cases Expanded

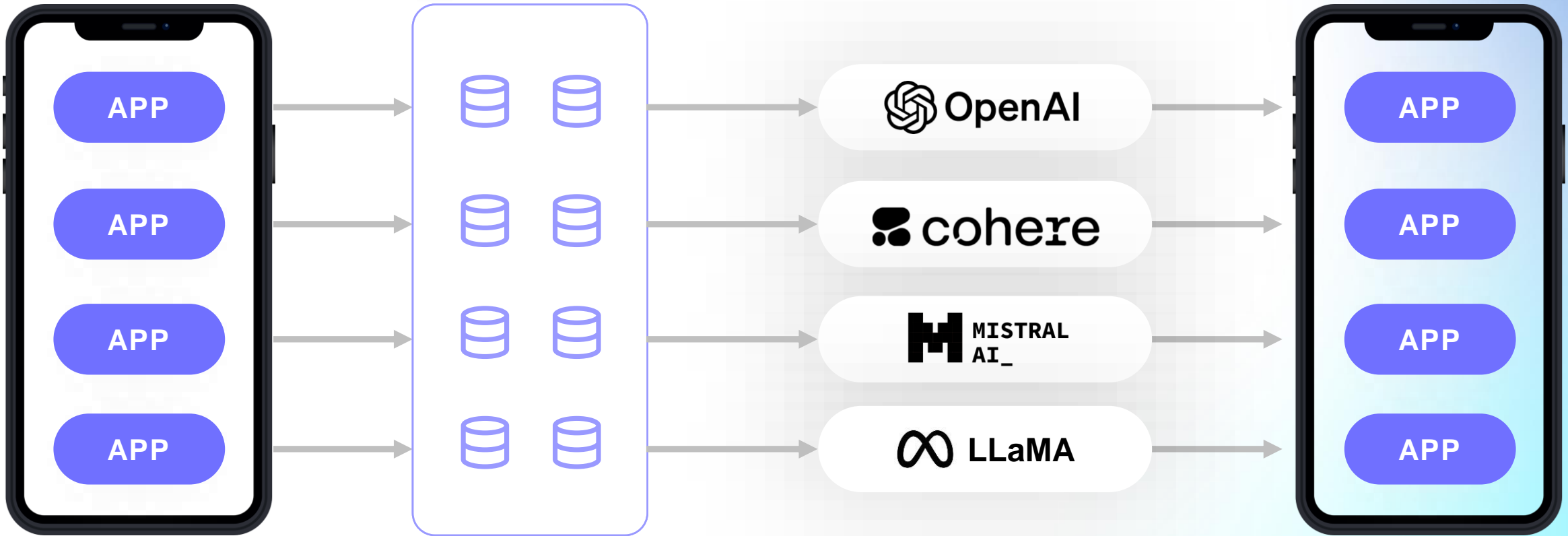


Across Multiple LLMs



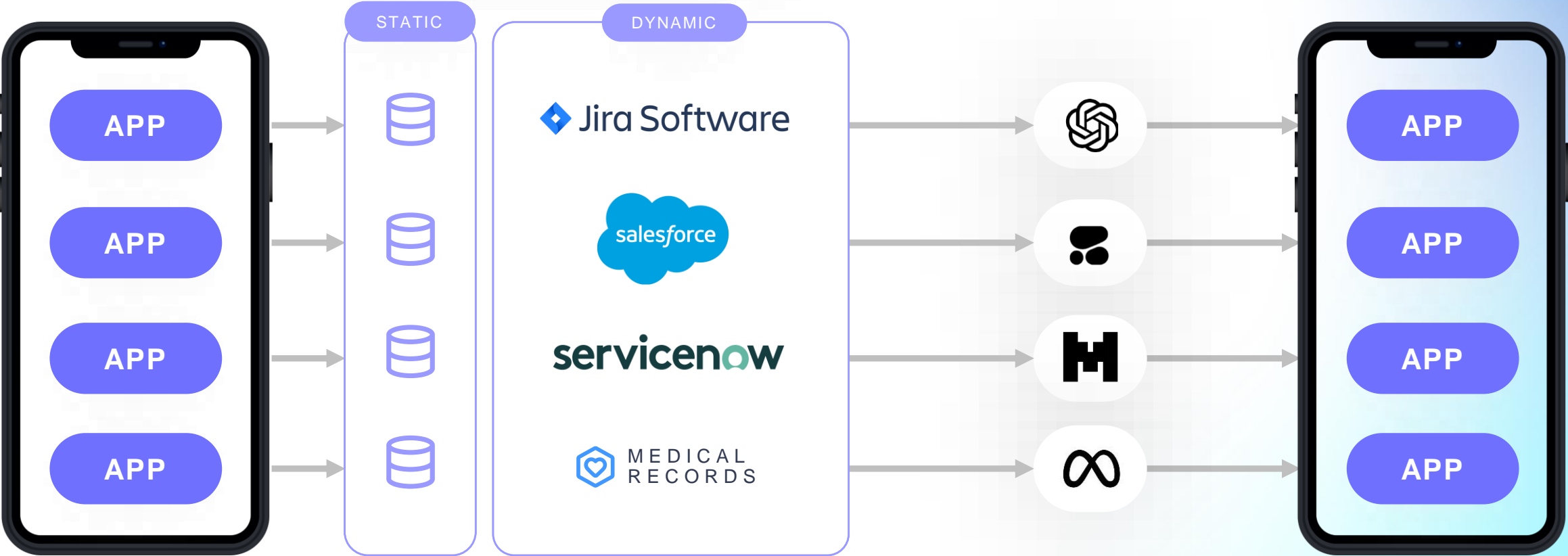
Combined with Company Information

ADDITIONAL DATA SOURCES

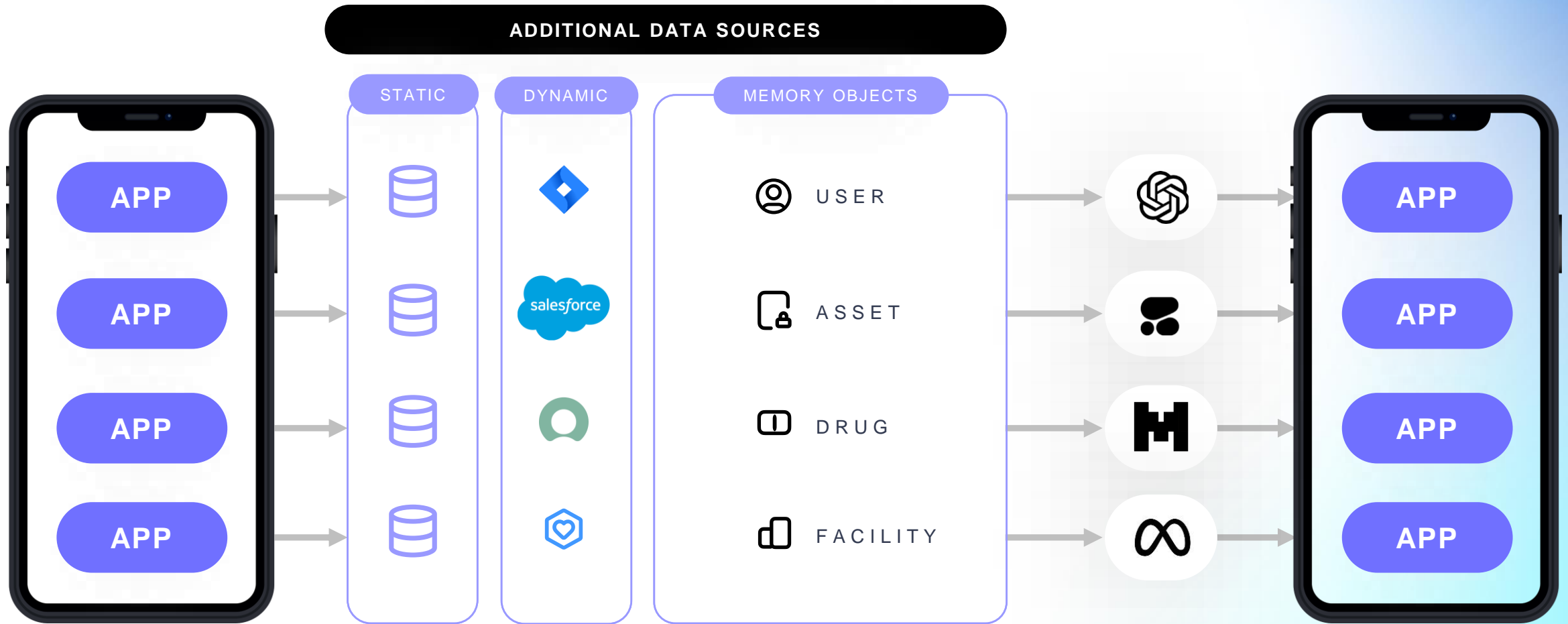


Transactional Data

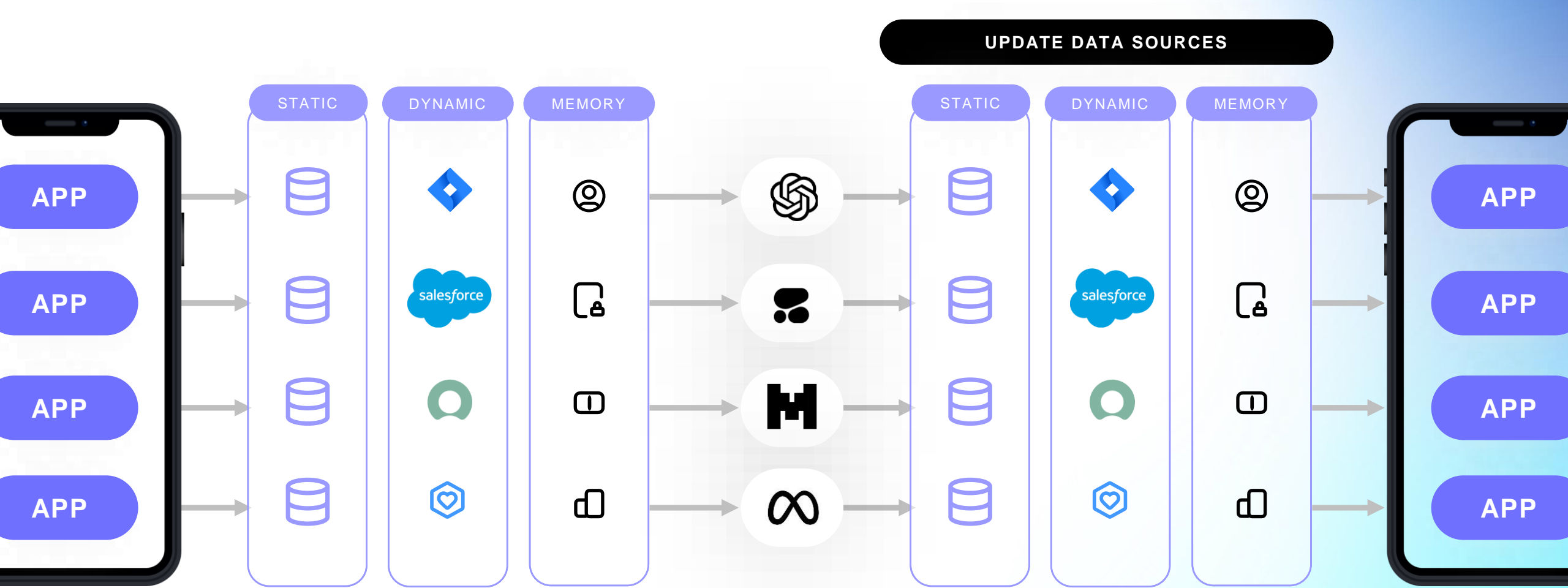
ADDITIONAL DATA SOURCES



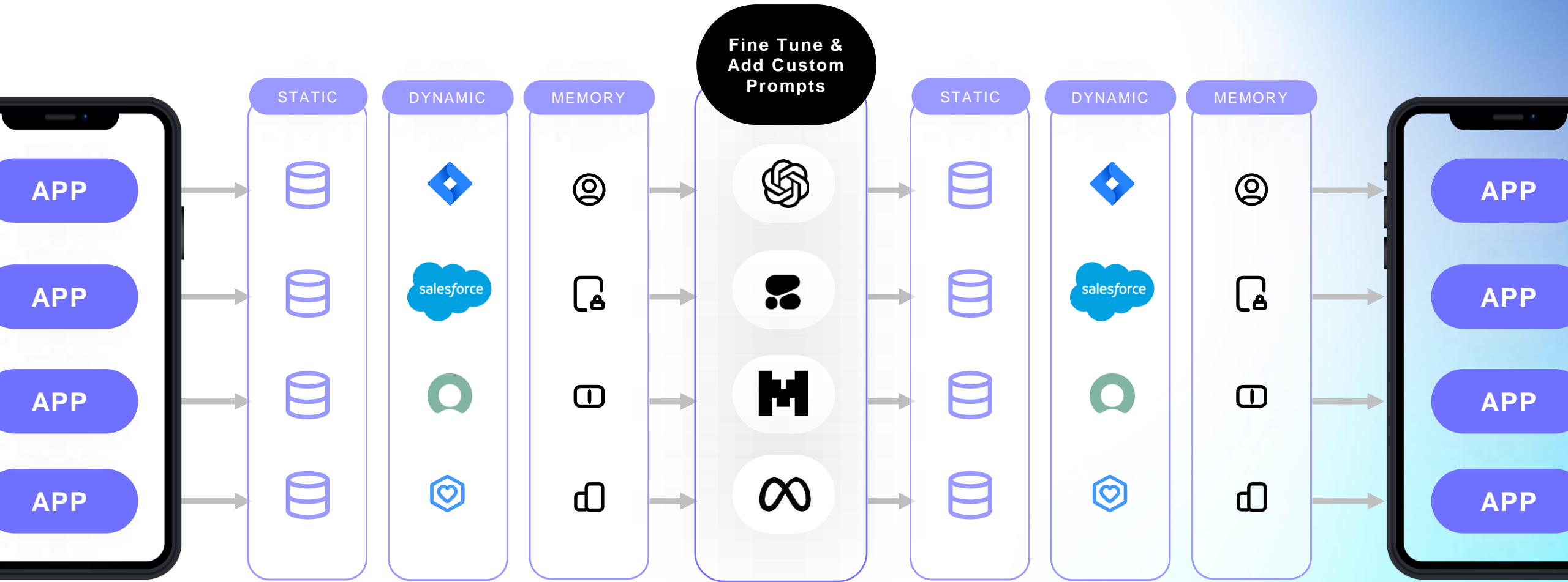
Plus Memory Objects



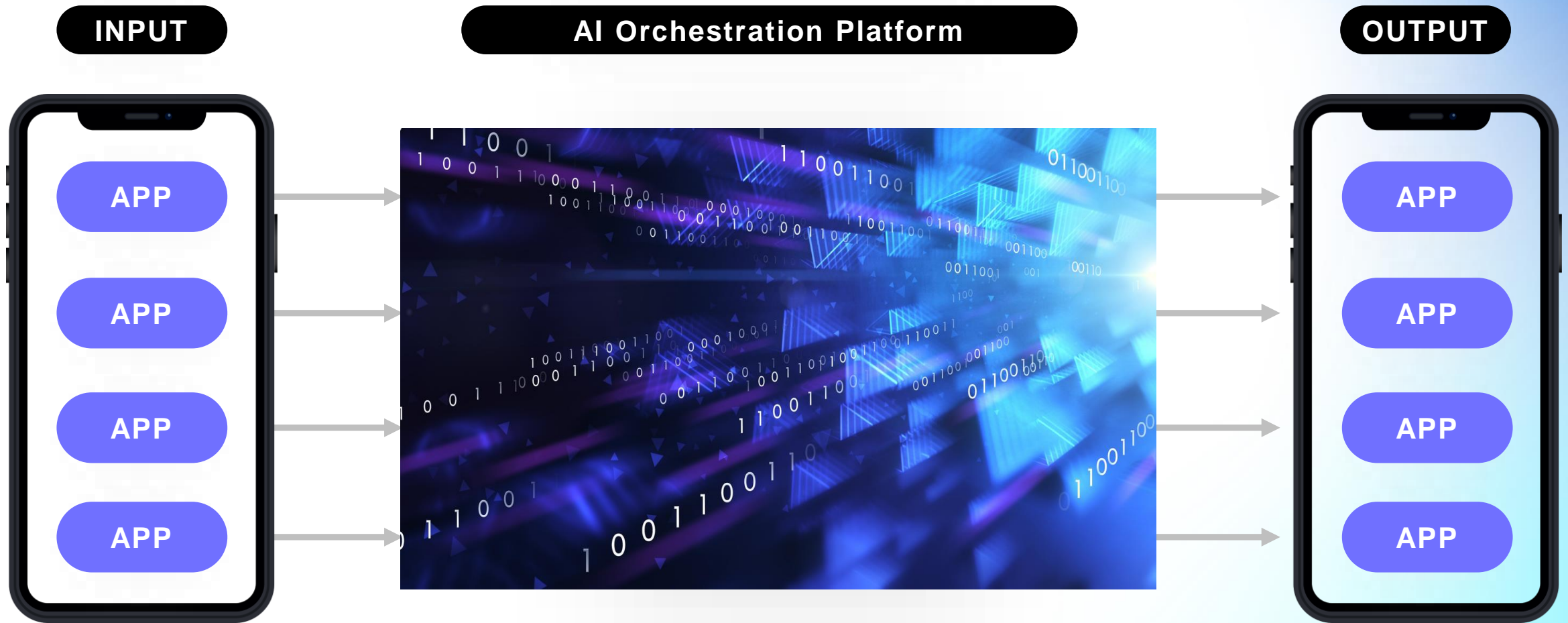
Write Updates to Data Sources



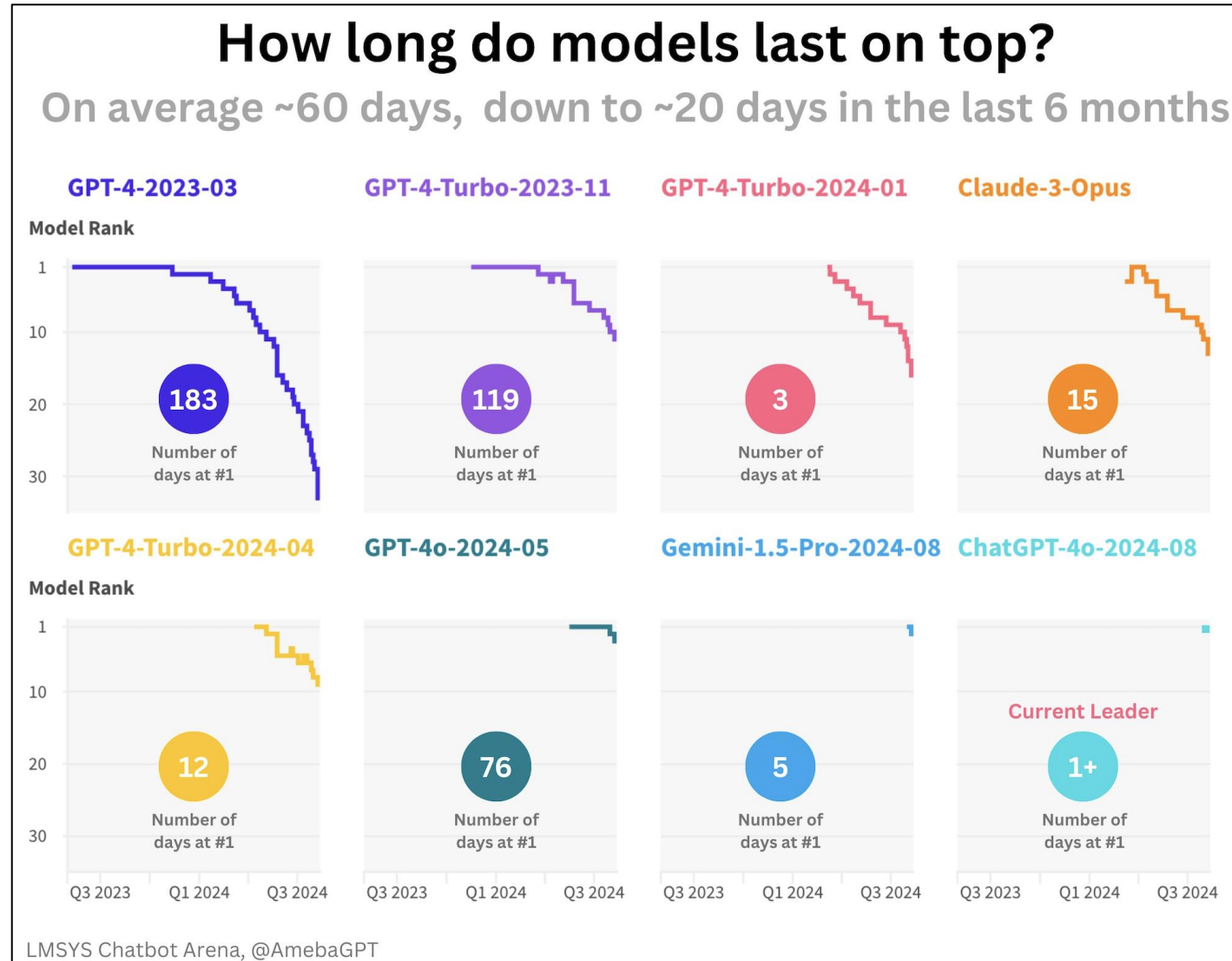
Maximize Value with Prompts & Fine Tuning



Abstract Operational Complexity for Success



Fast-Changing Leaderboard



Rapid Deprecation of “Old” Models

Subject: Final Deprecation Reminder: gpt-3.5-turbo-0301, gpt-3.5-turbo-0613, and gpt-3.5-turbo-16k-0613



This is a final reminder that the following models will **no longer be available** starting next Friday, September 13, 2024:

- **gpt-3.5-turbo-0301**
- **gpt-3.5-turbo-0613**
- **gpt-3.5-turbo-16k-0613**

We have noticed that your organization has recently used at least one of these models. To avoid any disruption, we encourage you to migrate to **gpt-4o-mini**, our new small model that will give you higher performance at lower cost.

Thanks for building with OpenAI. If you have any questions about model migration, feel free to reach out on the [OpenAI Developer Forum](#).

—The OpenAI Team

Subject: Deprecation Reminder: GPT-4 Vision Preview will be shut down on December 6, 2024



In **June 2024** we announced the deprecation of the following models, with a planned shutdown date of December 6, 2024:

- **gpt-4-vision-preview**
- **gpt-4-1106-vision-preview**

We have noticed that your organization has recently used one of these models. We encourage you to migrate to **GPT-4o**, our latest flagship model, as a replacement.

Thanks for building with OpenAI. If you have any questions about model migration, feel free to reach out on the [OpenAI Developer Forum](#).

—The OpenAI Team

Significant Cost Differences

Same prompt, same model provider, similar responses but wildly different costs

The image displays three sequential workflow panels, each representing a different model configuration. Each panel includes an 'Input' node, a 'Model' node, and an 'Output' node. The 'Model' nodes specify the model name, temperature, system prompt, and the input prompt. The 'Output' nodes show the assistant's response, the number of tokens generated, the time taken, and the cost per 100 calls. The costs are highlighted in blue callouts.

Model	Cost / 100 calls
gpt-4o	\$0.95
gpt-4o-2024-08-06	\$0.51
gpt-4o-mini	\$0.02

Panel 1: gpt-4o
Model: gpt-4o, Temperature: 0.7, System Prompt: ---, Prompt: ---
Assistant: 694 tokens, 8,422ms
Output: When comparing the cost and quality of OpenAI models, it is important to use benchmarks that comprehensively evaluate different aspects of model performance. assess these dimensions:
GLUE Benchmark (General Language Understanding Evaluation):
Description: GLUE is a collection of nine natural language understanding tasks, including tasks like sentiment analysis, text similarity, and textual entailment.
Cost: \$0.95 / 100 calls

Panel 2: gpt-4o-2024-08-06
Model: gpt-4o-2024-08-06, Temperature: 0.7, System Prompt: ---, Prompt: ---
Assistant: 546 tokens, 6,665ms
Output: When comparing the cost and quality of OpenAI models, it's important to consider a variety of benchmarks that assess different aspects of performance, including capabilities. Here are some widely recognized benchmarks and methods that can help you evaluate OpenAI models:
GLUE (General Language Understanding Evaluation): This benchmark suite is useful for evaluating natural language understanding tasks. It includes tasks like sentiment analysis, sentence similarity, and more. It provides a good measure of a model's capability to understand and process language.
Cost: \$0.51 / 100 calls

Panel 3: gpt-4o-mini
Model: gpt-4o-mini, Temperature: 0.7, System Prompt: ---, Prompt: ---
Assistant: 464 tokens, 4,809ms
Output: When comparing the cost and quality of OpenAI models, several benchmarks and metrics can be useful. Here are some key aspects to consider:
Performance Benchmarks:
GLUE/SuperGLUE: These are collections of natural language understanding tasks that test various aspects of language models, including sentiment analysis, textual entailment, and more.
Cost: \$0.02 / 100 calls

Entirely New Security Threat Vectors

New attack surfaces across the AI lifecycle



- Training data poisoning
- Data theft
- Untrusted inference infrastructure
- Model theft
- Uncertain model provenance


- Shadow AI
- Data leakage
- Model denial-of-service attacks
- Prompt injections
- Insecure tool calls

AI Brand Risk

Significant questions about how AI could impact public perception



Biased Algorithms




Exposing PII



Low Quality Output



Insensitive Information

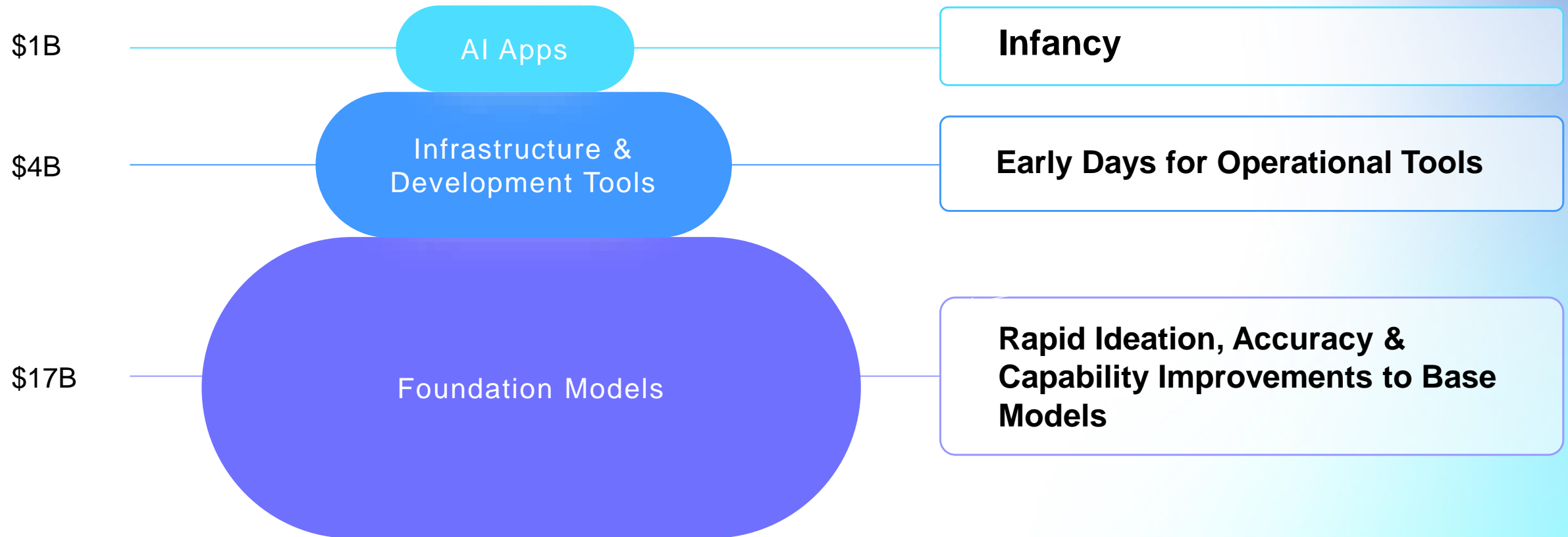


Misinformation in Training Data



Hallucinations

AI VC Investments To Date





Your company data has always
been one of your most valuable
assets.

Now, it's not just your data, it's

your AI

- Data Integration
- Models
- Prompts
- Tuning
- Policies
- Governance

Now, it's not just your data, it's
your AI Strategy & Platform.

own AI. How do I avoid vendor lock in. What are AI Agents. What is computer vision? What is AGI/Artificial General Intelligence? What is the singularity? How smart is AI? How will AI affect our capital and operating expenditures? What is Jailbreaking an LLM? What is accelerated computing? How can an AI be made better? Why should I care about AI? What is parallelization? What is a

H100? Does OpenAI use my questions to train other models (or LLMs)? What is a private model? How do I use a private model? Is it safe to give an LLM my contracts and info? Can an LLM write documents for me? What causes a model to hallucinate? How do I trust answers from LLMs aren't hallucinations? Should I train a model for my company? How

**how can
airia help me
with AI?_>**

long does it take to train a model on my data? How do I give an LLM sensitive information from my company? Can AI be secure? What is the future of AI in my field? How can I use AI in my work? How reliable are AI models? What are the potential risks of error or bias? What happens when AI gets it wrong? How does AI



Airia is an enterprise AI, full-stack platform that quickly and securely modernizes workflows, deploys and manages industry-leading AI models, and provides instant time-to-value for impactful ROI.